Ulrike Kuhl
Bielefeld University, Bielefeld, Germany

Mark T. Keane
University College Dublin, Dublin, Ireland

Six Steps to Heaven…

Motivation

Design

Materials & Procedure

Piloting

Data Collection & Analysis

Results

To what end do you want to test what with whom in which context?

Motivation

To what end?

Examine how the **personal relevance** of AI system decisions impacts user behavior and perception of explanations.

What?

Human **prediction accuracy** w.r.t. system decisions that are:

- **either relevant for oneself or another person**, and
- either supported by **counterfactuals or control** descriptions

+ subjective satisfaction & trust in the corresponding system

Whom?

AI experts attending this tutorial @IJCAI 2024

To what end do you want to test what with whom in which context?

Motivation

Context?

AI-based performance evaluation and reward system *PERS*

Factor 1: **self-relevant** **vs.** **other-relevant**

Case 1/16

Please review the information and take a guess.

YOUR performance:

Productivity	Low
Quality of work	Low
Customer feedback	High
Punctuality	High
Bonus?	??

Will YOU get the bonus?

Case 1/16

Please review the information and take a guess.

Frank's performance:

Punctuality	High
Quality of work	Low
Productivity	High
Customer feedback	Low
Bonus?	??

Will Frank get the bonus?

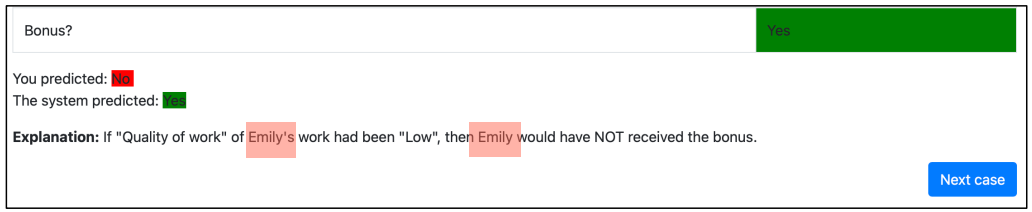
To what end do you want to test what with whom in which context?

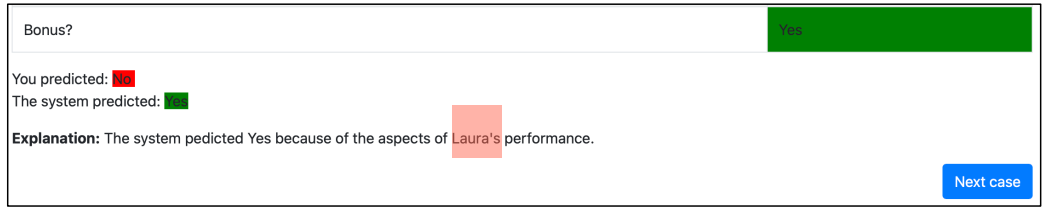
Motivation

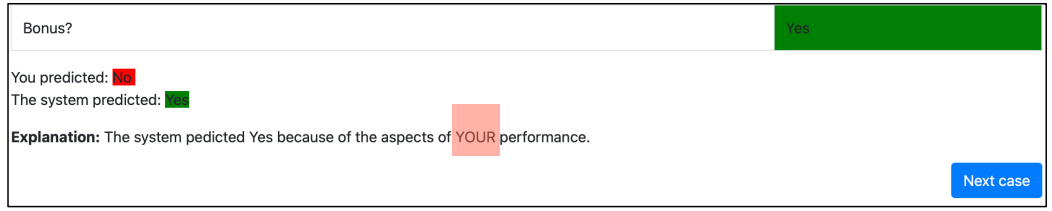
Context?

AI-based performance evaluation and reward system *PERS*

Factor 2: **counterfactual** **vs.** **control description**



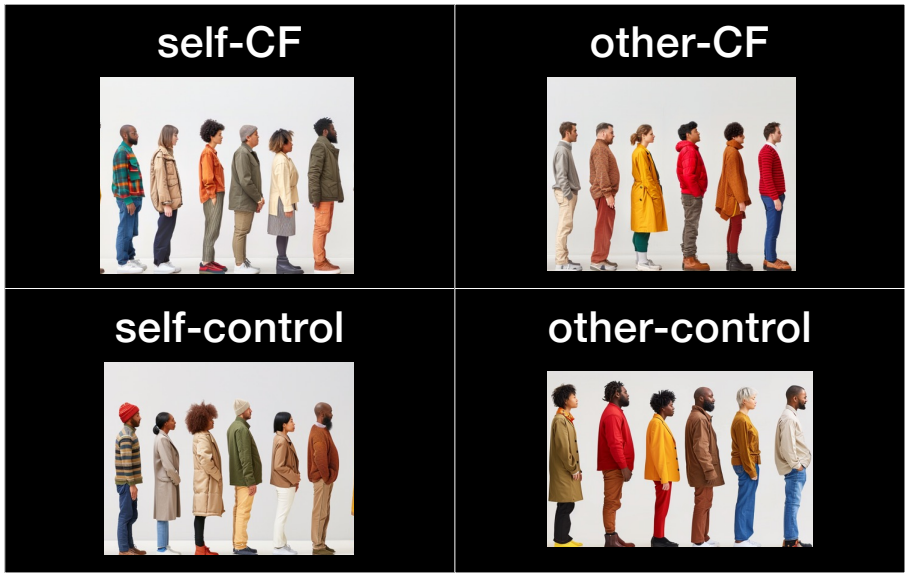




Design-E

2 x 2

STEP 2 Design



We used a ***between-participant*** design, with ***two factors*** (self-relevance, explanation), with **two levels** each:

self-relevance has levels self vs. other

explanation has levels counterfactual vs. control

Back to balancing and randomizing

Materials & Procedure

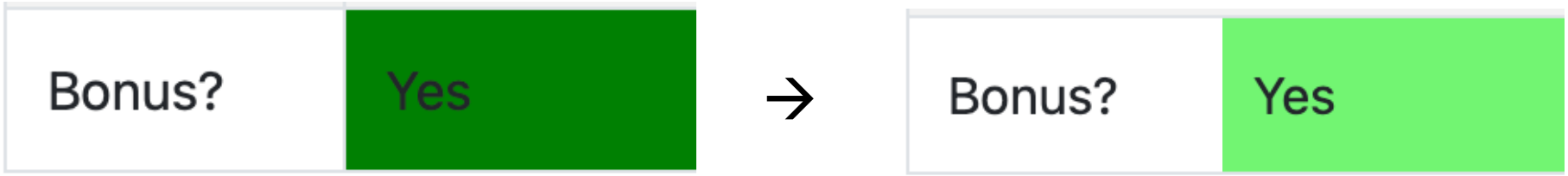
Quality of work	H	H	H	H	H	H	H	H	L	L	L	L	L	L	L	L
Productivity	H	H	H	H	L	L	L	L	H	H	H	H	L	L	L	L
Punctuality	H	H	L	L	H	H	L	L	H	H	L	L	H	H	L	L
Customer feedback	H	L	H	L	H	L	H	L	H	L	H	L	H	L	H	L
Bonus?	Y	Y	Y	Y	Y	Y	N	N	Y	Y	N	N	N	N	N	N

- ✓ ½ items are positive ½ items are negative
- ✓ presented order of features randomized across participants
- ✓ **presented order of items randomized across participants**
- ✓ other-relevant condition: ½ items shown as male gendered, ½ items shown as female gendered

Pilot: Done with family and friends beforehand

Piloting

N=37 → 1 failed BOTH attention checks → N=36



Instruction ”other” conditions:

Imagine SEVERAL EMPLOYEES THAT work…

→ Imagine SEVERAL EMPLOYEES that work…

```
Summary of the number of participants in each group:
    cond  Number of Participants
0  o-cfe                     16
1  o-con                      8
2  s-cfe                      7
3  s-con                      5
```

Randomization produced unbalanced groups!

Pilot: Done with family and friends beforehand

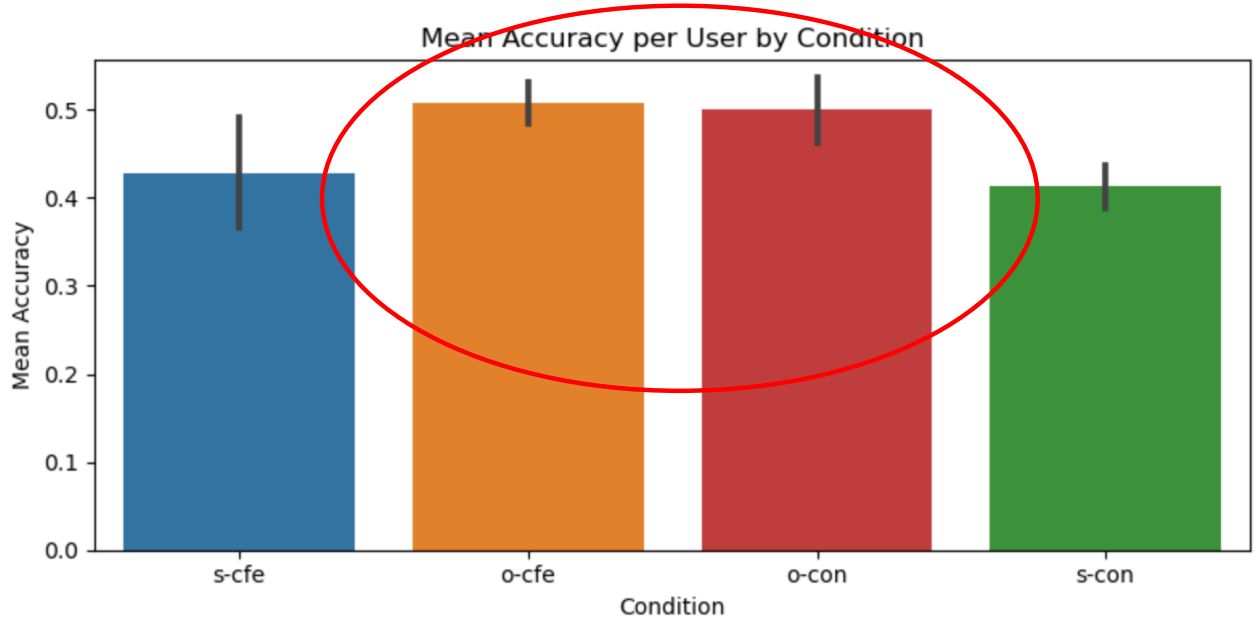
Piloting

Mean accuracy over task:

Punctuality	High
Bonus?	??

Will YOU get the bonus?

◯ Yes ◯ No



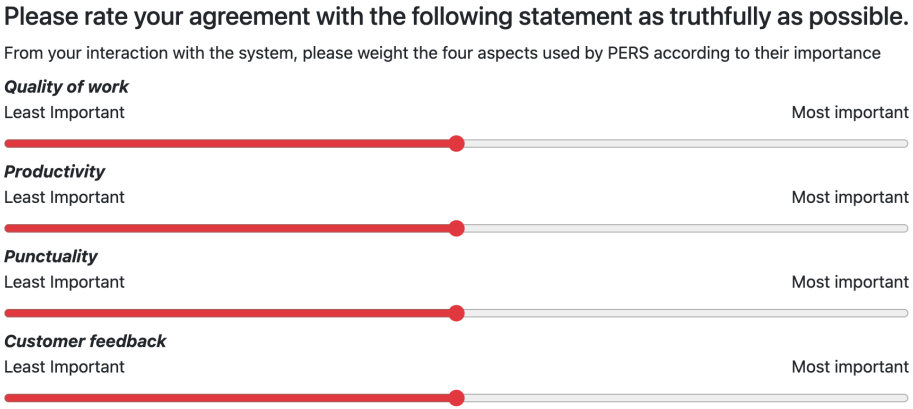
looks like ‘other’ > ‘self’! (ANOVA confirmed!)

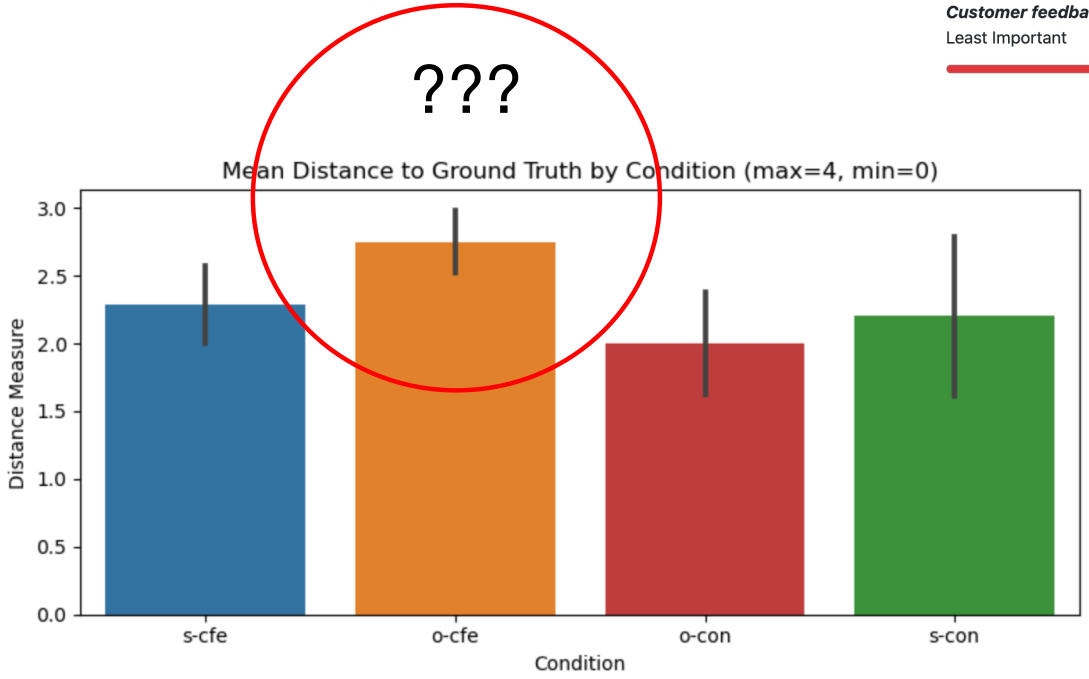
Pilot: Done with family and friends beforehand

Piloting

How close to ground truth?:

looks like CFE > CON

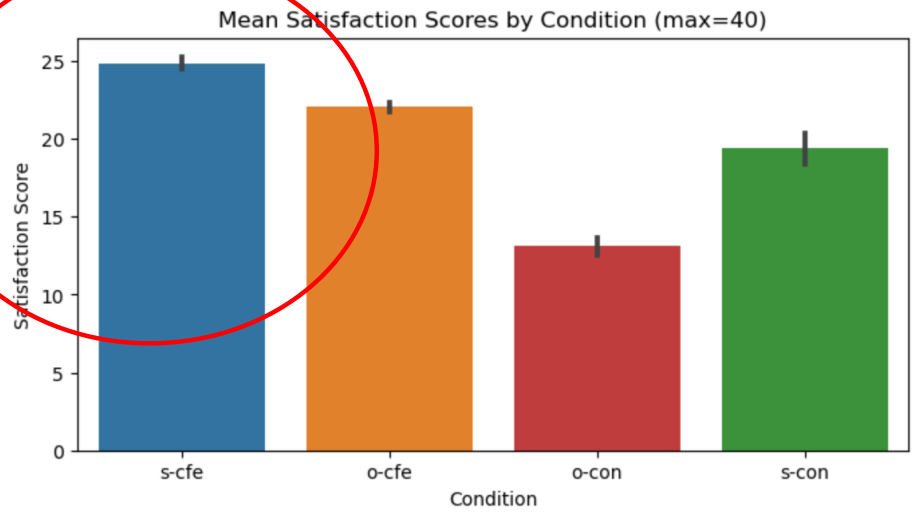




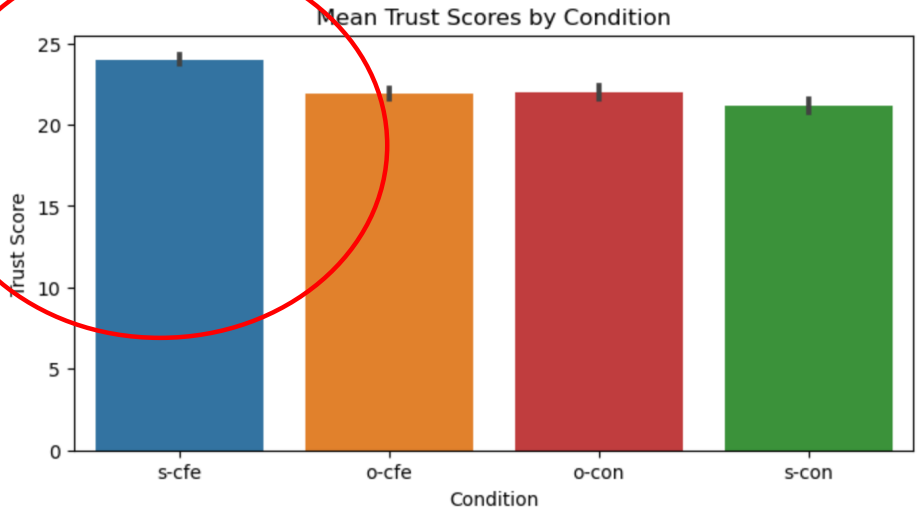
Piloting

Pilot: Done with family and friends beforehand

Satisfaction



Trust

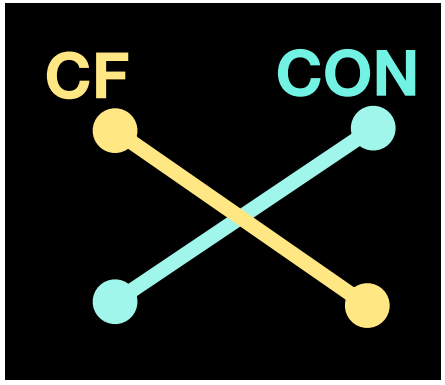


"s-cfe" higher than all others?

(Before) Starting the Statistics

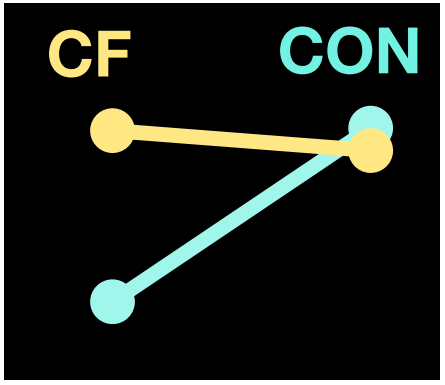
Data Collection & Analysis

→ A quick look at 2x2 designs, and how the means could behave



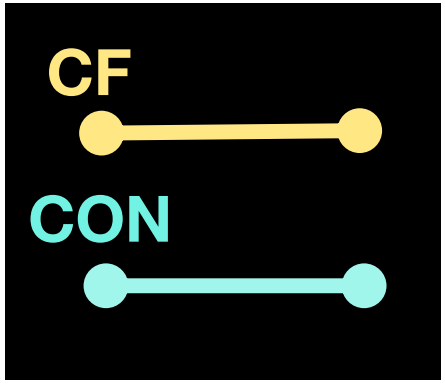
self other

strong interaction



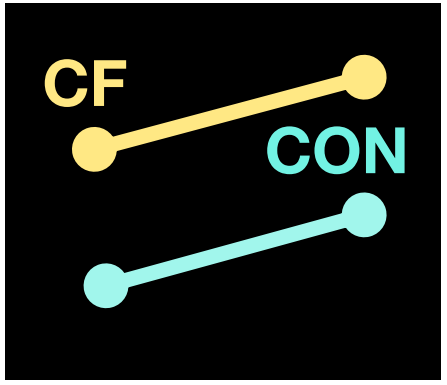
self other

weaker interaction



self other

Independent, no difference



self other

independent, with difference

Doing the Statistics – for real!

Data Collection & Analysis

Results

Let’s do it live!
https://colab.research.google.com/github/andreArtelt/IJCAI24-CF_Tut/blob/main/CESORP_study_evaluation.ipynb